

# Asymptotic separability in sensitivity analysis

Joseph L. Gastwirth

*George Washington University, Washington DC, USA*

and Abba M. Krieger and Paul R. Rosenbaum

*University of Pennsylvania, Philadelphia, USA*

[Received September 1998. Final revision November 1999]

**Summary.** In an observational study in which each treated subject is matched to several untreated controls by using observed pretreatment covariates, a sensitivity analysis asks how hidden biases due to unobserved covariates might alter the conclusions. The bounds required for a sensitivity analysis are the solution to an optimization problem. In general, this optimization problem is not separable, in the sense that one cannot find the needed optimum by performing a separate optimization in each matched set and combining the results. We show, however, that this optimization problem is asymptotically separable, so that when there are many matched sets a separate optimization may be performed in each matched set and the results combined to yield the correct optimum with negligible error. This is true when the Wilcoxon rank sum test or the Hodges–Lehmann aligned rank test is applied in matching with multiple controls. Numerical calculations show that the asymptotic approximation performs well with as few as 10 matched sets. In the case of the rank sum test, a table is given containing the separable solution. With this table, only simple arithmetic is required to conduct the sensitivity analysis. The method also supplies estimates, such as the Hodges–Lehmann estimate, and confidence intervals associated with rank tests. The method is illustrated in a study of dropping out of US high schools and the effects on cognitive test scores.

**Keywords:** Aligned rank test; Hodges–Lehmann estimate; Observational studies; Permutation inference; Randomization inference; Rank sum test; Sensitivity analysis

## 1. Sensitivity analysis in observational studies

In an observational study or non-randomized experiment, a sensitivity analysis asks how departures from random assignment of treatments of various magnitudes might alter the study's conclusions. It has been discussed by Cornfield *et al.* (1959), Bross (1966), Greenhouse (1982), Rosenbaum and Rubin (1983), Rosenbaum (1986, 1988, 1993, 1995), Rosenbaum and Krieger (1990), Gail *et al.* (1988), Gastwirth (1992), Angrist *et al.* (1996), Copas and Li (1997), Gastwirth and Nayak (1997) and Lin *et al.* (1998).

The sensitivity analysis consists of computing bounds on inferences for departures from random assignment of various magnitudes—i.e. bounds on significance levels, confidence intervals and point estimates. A study is sensitive to hidden bias if small departures from randomization can materially alter the study's conclusions, and it is insensitive if only large departures can affect the inferences. Cornfield *et al.* (1959) showed that for an unobserved attribute to explain the association between smoking and lung cancer it would need to predict lung cancer almost perfectly and be nine times more common among heavy smokers than among non-smokers.

*Address for correspondence:* Paul R. Rosenbaum, Department of Statistics, Wharton School, University of Pennsylvania, 3000 Steinberg Hall–Dietrich Hall, 3620 Locust Walk, Philadelphia, PA 19104-6302, USA.  
E-mail: rosenbaum@stat.wharton.upenn.edu

The potential for hidden bias is investigated by finding extreme values of the unobserved covariate vector  $\mathbf{u}$  of dimension  $N$  for the  $N$  subjects. For an estimate, we find the  $\mathbf{u}$  that produces the maximum estimate and the  $\mathbf{u}$  which produces the minimum estimate. Similarly, we find the  $\mathbf{u}$ s that maximize and minimize a  $P$ -value or the end point of a confidence interval.

For many cases, simple calculations yield the required sensitivity bounds. In some cases, the extreme  $\mathbf{u}$  can be identified immediately; for instance, this is true of methods for matched pairs, such as the signed rank statistic and the paired  $t$  test, and for methods with binary outcomes, such as Fisher's test for a  $2 \times 2$  table, the Mantel–Haenszel statistic and quantile procedures. In other cases, there is a small set of candidates for the extreme  $\mathbf{u}$  which may be quickly compared; for instance, this is true of methods for comparing two groups, such as the rank sum statistic, the Gehan and log-rank statistics and the two-sample  $t$  statistic (Rosenbaum (1995), chapter 4).

However, there are other cases where finding the extreme  $\mathbf{u}$  is difficult because the set of candidates is very large. One such situation, discussed in Section 2, is matching with multiple controls using the stratified Wilcoxon rank sum statistic. The difficulty is that the problem of finding the extreme  $\mathbf{u}$  is not separable—it cannot be done separately in each matched set. This paper introduces an asymptotically separable approximation to the solution of an inseparable optimization problem. An approximate  $\mathbf{u}$  is determined by optimizing separately each matched set. This easily computed  $\mathbf{u}$  yields an estimate or significance level differing from the true optimum by an amount that becomes arbitrarily small as the number of matched sets increases. In Section 3, we show how to approximate closely the bounds that would be obtained from the computational unattainable extreme  $\mathbf{u}$ . In Section 4, we state and prove the basic theorem justifying our separable approximation.

## 2. Matching with multiple controls

### 2.1. Randomization inference in experiments: tests, estimates and confidence intervals

There are  $I$  matched sets formed on the basis of observed pretreatment variables or covariates. In set  $i$ , one subject received the treatment and  $n_i - 1$  subjects received the control. Write  $Z_{ij} = 1$  if the  $j$ th subject in set  $i$  received the treatment and  $Z_{ij} = 0$  if this subject received the control, so  $1 = \sum_{j=1}^{n_i} Z_{ij}$  for each  $i$ . There are  $N = \sum n_i$  subjects in total. Write  $\mathbf{Z}$  for the  $N$ -dimensional vector containing the  $Z_{ij}$  in the lexical order,  $\mathbf{Z} = (Z_{11}, Z_{12}, \dots, Z_{1,n_1}, Z_{21}, \dots, Z_{I,n_I})^T$ . Let  $B$  be the set containing the  $M = \prod n_i$  possible values of  $\mathbf{Z}$ , i.e.  $B$  contains  $M$  vectors  $\mathbf{z} = (z_{11}, z_{12}, \dots, z_{I,n_I})^T$  of dimension  $N$  with 0 or 1 co-ordinates such that each vector  $\mathbf{z} \in B$  satisfies  $1 = \sum_{j=1}^{n_i} z_{ij}$  for  $i = 1, \dots, I$ . Treatment assignments in distinct matched sets are assumed to be independent.

A subject exhibits a response  $R_{ij}$ , and the treatment effect is assumed to be additive, so  $R_{ij} = r_{ij} + Z_{ij}\tau$ , where  $r_{ij}$  is the response that the  $j$ th subject in set  $i$  would exhibit under the control. The  $r_{ij}$  are fixed, not varying with the treatment assignment, whereas the observed responses  $R_{ij}$  are random variables that are affected by the treatment  $Z_{ij}$ —this is the traditional formulation of randomization inference (Fisher (1935), Kempthorne (1952) and Lehmann (1986), chapter 5). Write  $\mathbf{R} = (R_{11}, R_{12}, \dots, R_{I,n_I})^T$  and  $\mathbf{r}$  for the vector of  $r_{ij}$ , so  $\mathbf{r} = \mathbf{R} - \mathbf{Z}\tau$  is fixed.

A test of the hypothesis  $H_0: \tau = \tau_0$  will lead to point estimates and confidence intervals. Let  $\mathbf{q} = \mathbf{q}(\mathbf{R} - \mathbf{Z}\tau_0)$  be an  $N$ -dimensional function that assigns scores to the adjusted responses  $\mathbf{R} - \mathbf{Z}\tau_0$ , and consider the statistic  $T = \mathbf{Z}^T \mathbf{q}$ . A common choice for the function  $\mathbf{q}$  is to set  $q_{ij}$  equal to the rank of  $R_{ij} - Z_{ij}\tau_0$  among the  $n_i$  subjects in set  $i$ , with average ranks for ties, so that  $\mathbf{Z}^T \mathbf{q}$  is the stratified Wilcoxon rank sum statistic. A second common choice for  $\mathbf{q}$  involves

subtracting from each  $R_{ij} - Z_{ij}\tau_0$  the mean of the  $(R_{ij} - Z_{ij}\tau_0)$ s within set  $i$ , and ranking the results from 1 to  $N$ , so that  $\mathbf{Z}^T \mathbf{q}$  is the aligned rank statistic of Hodges and Lehmann (1962). Under  $H_0$ ,  $\mathbf{q} = \mathbf{q}(\mathbf{R} - \mathbf{Z}\tau_0)$  is fixed, and  $\mathbf{Z}^T \mathbf{q}$  is a sum of scores for treated subjects.

In a matched randomized experiment with multiple controls,  $\text{prob}(Z_{ij} = 1) = 1/n_i$  for each  $i$ , and each  $\mathbf{z} \in B$  has the same probability  $1/M$ . From this, the randomization distribution of  $T = \mathbf{Z}^T \mathbf{q}(\mathbf{R} - \mathbf{Z}\tau_0)$  is obtained under the null hypothesis  $H_0: \tau = \tau_0$ . As  $T$  is the sum of  $I$  independent random variables, its distribution may often be approximated as  $I \rightarrow \infty$  by a normal distribution.

The null randomization distribution of  $T$  is inverted to obtain confidence intervals for  $\tau$  as the set of all  $\tau$ s not rejected by the test. If, as is often true, the null expectation  $E_0$  of  $T$  does not depend on  $\tau$ , then solving  $E_0 = \mathbf{Z}^T \mathbf{q}(\mathbf{R} - \mathbf{Z}\hat{\tau})$  yields the Hodges and Lehmann (1963) estimate  $\hat{\tau}$  of  $\tau$ . In an experiment, if  $T$  is the stratified rank sum statistic, then  $E_0 = \Sigma (n_i + 1)/2$ .

2.2. Brief review of a model for sensitivity analysis in an observational study

A model for sensitivity analysis in observational studies is briefly reviewed in this section; see Rosenbaum (1988, 1995) and Rosenbaum and Krieger (1990) for details. The model assumes that subjects matched on the basis of observed covariates may differ in their chances of receiving the treatment because they differ with respect to unobserved covariates. Suppose that the  $Z_{ij}$  were initially independent with  $\text{prob}(Z_{ij} = 1) = p_{ij}$ , that the  $p_{ij}$  satisfied the inequality

$$\frac{1}{\Gamma} \leq \frac{p_{ij}(1 - p_{ik})}{p_{ik}(1 - p_{ij})} \leq \Gamma \quad \text{for } i = 1, \dots, I, \quad 1 \leq j, k \leq n_i \quad (1)$$

with  $\Gamma \geq 1$ , and that attention is returned to  $B$  by conditioning on  $1 = \Sigma_{j=1}^{n_i} Z_{ij}$  for each  $i$ . In this model,  $\Gamma$  is a sensitivity parameter describing the degree to which subjects in the same matched set may differ in their chances of receiving the treatment. If  $\Gamma = 1$ , then all subjects in the same matched set have the same probability of receiving the treatment, and the conditioning then assigns equal probabilities  $1/M$  to each  $\mathbf{z} \in B$ , as in a randomized experiment. If  $\Gamma = 2$ , then two subjects in the same matched set may differ by a factor of 2 in their odds of receiving the treatment. This model is identical with assuming

$$\text{prob}(\mathbf{Z} = \mathbf{z}) = \prod_{i=1}^I \frac{\exp\left(\gamma \sum_{j=1}^{n_i} z_{ij} u_{ij}\right)}{\sum_{k=1}^{n_i} \exp(\gamma u_{ik})} \quad \text{for } \mathbf{z} \in B,$$

$$\text{with } 0 \leq u_{ij} \leq 1 \text{ for } i = 1, \dots, I, \quad j = 1, \dots, n_i, \quad (2)$$

where  $\gamma = \log(\Gamma)$ ; see Rosenbaum (1995), chapter 4. In equation (2),  $u_{ij}$  may be thought of as an unobserved binary covariate not used in the matching process. Write  $\mathbf{u}$  for the  $N$ -dimensional vector of  $u_{ij}$ .

For the stratified Wilcoxon statistic, Rosenbaum and Krieger (1990), section 3, showed that, after sorting the responses, the exact significance level  $\text{prob}(T \geq t)$  under model (2) is maximized at a vector  $\mathbf{u}$  with  $u_{i1} = u_{i2} = \dots = u_{i,a_i} = 0$  and  $u_{i,a_i+1} = \dots = u_{i,n_i} = 1$  for some  $a_i, i = 1, \dots, I$ . So there are  $\Pi (n_i - 1)$  candidate values of  $\mathbf{u}$  to be considered, or  $2^{100}$  candidates with  $I = 100$  matched triples,  $n_i = 3$ .

The problem can be approximated by using the asymptotic normality of  $T$  as  $I \rightarrow \infty$ . Since the asymptotic mean of  $T$  increases with order  $I$  and the standard deviation with order  $\sqrt{I}$ ,

the algorithm finds the  $u_{ij}$  in each stratum separately that maximizes the contribution to the asymptotic mean of  $T$ . When multiple  $u_{ij}$ s achieve the maximum mean, the  $u_{ij}$  maximizing the variance is chosen.

### 3. A separable algorithm

#### 3.1. Largest expectation and largest variance among tied expectations

The goal is an approximation to the upper tail area  $\text{prob}(T \geq t)$  under model (2) of the test statistic  $T = \mathbf{Z}^T \mathbf{q}$  under the null hypothesis  $H_0: \tau = \tau_0$  so  $\mathbf{q} = \mathbf{q}(\mathbf{R} - \mathbf{Z}\tau_0)$  is fixed. Renumber the  $n_i$  subjects in stratum  $i$  so that  $q_{i1} \leq q_{i2} \leq \dots \leq q_{i,n_i}$ . Let  $\Gamma = \exp(\gamma) \geq 1$ . When  $u_{i1} = u_{i2} = \dots = u_{i,a} = 0$  and  $u_{i,a+1} = \dots = u_{i,n_i} = 1$ , matched set  $i$  contributes  $\sum_{j=1}^{n_i} Z_{ij}q_{ij}$  with expectation and variance

$$\mu_{ia} = \frac{\sum_{j=1}^a q_{ij} + \Gamma \sum_{j=a+1}^{n_i} q_{ij}}{a + \Gamma(n_i - a)}$$

and

$$\nu_{ia}^2 = \frac{\sum_{j=1}^a q_{ij}^2 + \Gamma \sum_{j=a+1}^{n_i} q_{ij}^2}{a + \Gamma(n_i - a)} - \mu_{ia}^2.$$

For  $i = 1, \dots, I$ , define

$$\mu_i = \max_{a \in \{1, \dots, n_i - 1\}} (\mu_{ia}). \tag{3}$$

Then  $\mu_i$  is the maximum expectation of  $\sum_{j=1}^{n_i} Z_{ij}q_{ij}$  under model (2). It may happen that there is a unique  $a \in \{1, \dots, n_i - 1\}$  which maximizes equation (3) or it may happen that several values of  $a$  produce the same maximized value  $\mu_i$  for equation (3). In either case, let  $A \subseteq \{1, \dots, n_i - 1\}$  be the set of values of  $a$  which maximize equation (3). Then define

$$\nu_i^2 = \max_{a \in A} (\nu_{ia}^2), \tag{4}$$

so that  $\nu_i^2$  is the maximum variance of  $\sum_{j=1}^{n_i} Z_{ij}q_{ij}$  under model (2) among values of  $(u_{i1}, \dots, u_{i,n_i})$  that attain the maximum expectation  $\mu_i$ . Set  $\sigma_I^2 = \sum_{i=1}^I \nu_i^2$ .

If  $t > \sum_{i=1}^I \mu_i$ , then the approximate upper bound on the upper tail  $\text{prob}(T \geq t)$  is

$$1 - \Phi \left\{ \left( t - \sum_{i=1}^I \mu_i \right) / \sigma_I \right\}.$$

This is a large sample approximation to the actual bound.

#### 3.2. Stratified rank sum test and aligned rank test in matching with multiple controls

A simple, but practically important, case, in which the separable calculations of Section 3.1 work, is the stratified Wilcoxon rank sum test applied to matching with multiple controls. Indeed, the required calculations will be tabulated leaving only a small amount of arithmetic to be done. In testing  $H_0: \tau = \tau_0$  with this test, the  $n_i$  adjusted responses in set  $i$ ,  $R_{ij} - \tau_0 Z_{ij}$  are ranked from 1 to  $n_i$ , and  $T = \mathbf{Z}^T \mathbf{q}$  is the sum over matched sets of the  $I$  ranks for the  $I$  treated subjects.

**Table 1.** Separable rank sum calculations†

$n_i$	$\Gamma$	$\mu_i$	$\nu_i^2$	$n_i$	$\Gamma$	$\mu_i$	$\nu_i^2$
2	1	$\frac{3}{2}$	$\frac{1}{4}$	4	1	$\frac{5}{2}$	$\frac{5}{4}$
	2	$\frac{5}{3}$	$\frac{2}{9}$		2	$\frac{17}{6}$	$\frac{41}{36}$
	3	$\frac{7}{4}$	$\frac{3}{16}$		3	3	$\frac{4}{3}$
	4	$\frac{9}{5}$	$\frac{4}{25}$		4	$\frac{22}{7}$	$\frac{62}{49}$
3	1	2	$\frac{2}{3}$	5	1	3	2
	2	$\frac{9}{4}$	$\frac{11}{16}$		2	$\frac{24}{7}$	$\frac{96}{49}$
	3	$\frac{12}{5}$	$\frac{16}{25}$		3	$\frac{11}{3}$	$\frac{16}{9}$
	4	$\frac{5}{2}$	$\frac{7}{12}$		4	$\frac{42}{11}$	$\frac{194}{121}$

†In matched sets with one treated subject,  $n_i - 1$  controls and fixed  $\Gamma$ ,  $\mu_i$  is the maximum expectation of the rank sum and  $\nu_i^2$  is the maximum variance attainable at this maximum expectation.

Table 1 gives the required moments in equations (3) and (4) for matched sets with  $n = 2, 3, 4, 5$  subjects for values of  $\Gamma = 1, 2, 3, 4$ . To use Table 1,

- (a) calculate the stratified rank sum statistic  $T = \mathbf{Z}^T \mathbf{q}$ ,
- (b) fix a  $\Gamma$ ,
- (c) for matched set  $i, i = 1, \dots, I$ , use  $n_i$  in Table 1 to read  $\mu_i$  and  $\nu_i^2$  which are equal to expressions (3) and (4) respectively,
- (d) if  $T \leq \sum_{i=1}^I \mu_i$ , then the upper bound on the significance level is at least  $\frac{1}{2}$  and
- (e) otherwise, if  $T > \sum_{i=1}^I \mu_i$ , then approximate the upper bound on the significance level by

$$1 - \Phi \left\{ \left( T - \sum_{i=1}^I \mu_i \right) / \sqrt{\sum_{i=1}^I \nu_i^2} \right\}.$$

For the rank sum, a tabulation of the results of problems (3) and (4) is possible because the ranks in stratum  $i$  are always  $1, \dots, n_i$ . If there are ties, formulae (3) and (4) are used with average ranks.

An alternative to the stratified rank sum test is the aligned rank test (Hodges and Lehmann, 1962), which allows observations in different sets to be compared. In alignment the mean of the adjusted responses  $R_{ij} - \tau_0 Z_{ij}$  in matched set  $i$  is subtracted from each adjusted response in matched set  $i$ , and then some form of ranks are assigned across rather than within strata. See Lehmann (1975), Sen (1968) and Tardif (1981) for discussion of aligned ranks. The approximation uses the straightforward algorithm in Section 3.1. Now, however, a table such as Table 1 is not feasible and equations (3) and (4) are used directly. Analogous computations may be applied with full matching, the form of an optimal stratification (Rosenbaum, 1991; Gu and Rosenbaum, 1993).

### 3.3. A small numerical comparison: separable versus joint optimization

This section assesses the performance of the separable optimization, which is based on large sample theory, for the Wilcoxon rank sum test with  $I = 10$  or  $I = 15$  matched sets. Eight cases are considered in a  $2 \times 2 \times 2$  factorial formed from

- (a)  $I = 10$  or  $I = 15$  matched sets,
- (b)  $n_i = 3$  subjects per matched set, for  $i = 1, \dots, I$ , or  $n_i = 5$  subjects per matched set for  $i = 1, \dots, I$ , and
- (c)  $\Gamma = \exp(\gamma) = 2$  or  $\Gamma = 4$ .

The stratified Wilcoxon rank sum test is used. We want to approximate  $\text{prob}(T \geq k)$  for an interesting value of  $k$ . Somewhat arbitrarily, we picked  $k$  to be the value of the rank sum test corresponding to a standardized deviate of 3 in a randomized experiment, i.e. when  $\gamma = 0$ . In other words, we are examining the sensitivity of a result that would have a one-sided significance level close to 0.001 in a randomized experiment. Specifically,

$$k = I(n_i + 1)/2 + 3\sqrt{\{I(n_i - 1)(n_i + 1)/12\}},$$

for  $I = 10$  or  $I = 15$  and  $n_i = 3$  or  $n_i = 5$ .

Table 2 compares the minimum standardized deviate obtained by the method in Rosenbaum and Krieger (1990), section 5, labelled ‘ $m$ ’, to the separable approximation in this paper, labelled ‘ $s$ ’. The deviate would be referred to the normal distribution to approximate the bound on  $\text{prob}(T \geq k)$ . In seven of the eight cases, the results are identical, and in one case the results are close, disagreeing in the second decimal place.

The one case in Table 2 in which the results are not identical is  $I = 10, n_i = 5, \Gamma = 4$ . In that one case, the optimum choice of  $\mathbf{u}$  has  $(u_{i1}, \dots, u_{i5}) = (0, 0, 0, 0, 1)$  for  $i = 1, \dots, 10$  whereas the separable approximation has  $(u_{i1}, \dots, u_{i5}) = (0, 0, 0, 1, 1)$  for  $i = 1, \dots, 10$ . At the optimum choice,  $T$  has expectation 37.5 and variance 21.875, whereas at the separable choice, which maximizes the expectation,  $T$  has a larger expectation of 38.182 but a smaller variance of 16.033. Since the approximation works as  $I \rightarrow \infty$ , it is of interest to compare the results for  $I = 10, n_i = 5, \Gamma = 4$ , with the results for  $I = 15, n = 5, \Gamma = 4$ . In the latter case with  $I = 15$ , both the optimum and the separable choices of  $\mathbf{u}$  are the same, namely  $(u_{i1}, \dots, u_{i5}) = (0, 0, 0, 1, 1)$  for  $i = 1, \dots, 15$ . The separable algorithm picks  $(u_{i1}, \dots, u_{i5}) = (0, 0, 0, 1, 1)$  for each  $i$  and does not change as  $I$  increases. In contrast, as  $I$  increases, the optimum procedure switches to agree with the separable algorithm.

The  $I$  matched sets that form the basis for Table 2 are interchangeable with respect to  $i = 1, \dots, I$  because every matched set contains the same number  $n_i$  of subjects and the ranks in each set are the same, namely  $1, 2, \dots, n_i$ . In such a problem, the separable algorithm always picks the same  $(u_{i1}, \dots, u_{i5})$  for all  $i$ . In contrast, the optimum procedure may allow  $(u_{i1}, \dots, u_{i5})$  to vary with  $i$ . For instance, for  $\Gamma = 3.5, n_i = 5, i = 1, \dots, I = 10$ , the optimum  $\mathbf{u}$  has  $(u_{i1}, \dots, u_{i5}) = (0, 0, 0, 0, 1)$  for any eight  $i$ s and  $(u_{i1}, \dots, u_{i5}) = (0, 0, 0, 1, 1)$  for the remaining two  $i$ s. Of course, it does not matter which of the two strata have  $(u_{i1}, \dots, u_{i5}) = (0, 0, 0, 1, 1)$ , so there are  $\binom{10}{2} = 45$  equivalent optimal solutions. The optimum deviate is 1.431 and the separable approximation is 1.440. Here also, even though the optimum  $\mathbf{u}$  differs from the separable solution, the separable solution is a close approximation. These calculations suggest that the separable algorithm is appropriate even for sensitivity analyses of matched data sets of modest size.

**Table 2.** Comparison of minimum deviates and separable approximations†

$I$		$n_i = 3$		$n_i = 5$	
		$\Gamma = 2$	$\Gamma = 4$	$\Gamma = 2$	$\Gamma = 4$
10	$m$	2.001	1.137	2.063	1.265
	$s$	2.001	1.137	2.063	1.307
15	$m$	1.786	0.672	1.842	0.844
	$s$	1.786	0.672	1.842	0.844

† $I$  is the number of matched sets,  $n_i$  is the number of observations in set  $i$ ,  $i = 1, \dots, I$ ,  $m$  is the minimum deviate and  $s$  is the separable approximation.

### 3.4. An example: dropping out of high school and cognitive achievement

The example concerns the cognitive achievement of US high school drop-outs compared with that of similar students who remained in school (Rosenbaum, 1986). Students in the national sample High School and Beyond were tested in their sophomore year, 1980, before the drop-outs left school, and again in their senior year, 1982. Drop-outs were matched to students from the same school by using an estimate of the propensity to drop out based on 32 covariates describing students in their sophomore year, including the sophomore year test scores. To illustrate the calculations, we examine 12 drop-outs, each matched to two controls. (The original study had 2166 drop-outs and reached different conclusions because of its larger sample size.)

Table 3 gives test score declines from 1980 to 1982 for the 12 drop-outs and their controls. A negative score decline is, of course, a score gain. For instance, in matched set 1, the drop-out had

**Table 3.** US high school drop-outs matched to two controls

Set <i>i</i>	Group†	Score decline	Aligned values	Aligned rank	Expectation			Variance
					(0, 0, 1)	(0, 1, 1)	Maximum	
1	D	-11.39	1.75	21	19.50	19.80	19.80	30.96
	C	-8.45	4.69	24				
	C	-19.57	-6.43	9				
2	D	-8.33	0.07	18	21.75	21.00	21.75	126.19
	C	3.06	11.46	32				
	C	-19.93	-11.53	5				
3	D	-18.73	-5.97	10	20.00	20.00	20.00	37.50
	C	-11.99	0.77	20				
	C	-7.55	5.21	25				
4	D	11.94	13.21	33	24.50	25.80	25.80	154.56
	C	9.4	10.67	31				
	C	-25.16	-23.89	1				
5	D	-0.77	5.40	26	19.75	19.00	19.75	42.19
	C	-10.46	-4.29	11				
	C	-7.27	-1.10	16				
6	D	24.03	17.87	36	21.75	19.80	21.75	213.19
	C	-8.23	-14.39	3				
	C	2.67	-3.49	12				
7	D	-4.04	-0.10	17	18.50	18.20	18.50	14.25
	C	-6.67	-2.73	13				
	C	-1.12	2.82	22				
8	D	-14.4	-2.53	14	22.00	20.40	22.00	181.50
	C	-26.21	-14.34	4				
	C	5	16.87	35				
9	D	-1.7	6.54	28	20.75	20.40	20.75	67.69
	C	-15.3	-7.06	8				
	C	-7.73	0.51	19				
10	D	-0.87	10.22	30	20.50	19.40	20.50	98.25
	C	-19.71	-8.62	7				
	C	-12.69	-1.60	15				
11	D	-3.36	13.44	34	24.25	24.80	24.80	139.76
	C	-11.21	5.59	27				
	C	-35.83	-19.03	2				
12	D	5.89	6.86	29	21.75	22.00	22.00	71.20
	C	-10.79	-9.82	6				
	C	2.00	2.97	23				
Total							257.40	1177.23

†D, drop-out; C, control.

**Table 4.** Sensitivity analysis for the drop-out data

$\Gamma$	Range of significance levels	Range of Hodges–Lehmann estimates
1	[0.019, 0.019]	[8.16, 8.16]
1.35	[0.006, 0.050]	[6.50, 9.69]
2	[0.0009, 0.129]	[4.03, 11.86]
3	[0.0001, 0.278]	[1.86, 14.41]

a test score gain of 11.39 points from sophomore to senior year, which fell between the score gains for the two controls. Following Hodges and Lehmann (1962), test scores within each matched set were aligned by subtracting the mean in the matched set, and the aligned test scores were ranked from 1 to 36. The drop-out in matched set  $i = 6$  received a rank of 36.

The aligned rank statistic is the sum of the ranks for the 12 drop-outs, namely  $T = 296$ . In a randomized experiment, under the null hypothesis of no treatment effect,  $T$  has expectation  $(1 + 2 + \dots + 36)/3 = I(I + 1)/6 = 222$  and variance 1271, yielding a standardized deviate of  $(296 - 222)/\sqrt{1271} = 2.07$  and an approximate one-sided significance level of 0.019. If 8.16 is subtracted from the test score of each drop-out before computing the aligned rank statistic, then the statistic  $T$  equals its null expectation, 222, so in a randomized experiment the Hodges and Lehmann (1963) estimate of additive decline due to dropping out would be 8.16 points.

Table 4 gives the sensitivity calculations for  $\Gamma = 2$  based on the algorithm in Section 3.1. For  $\Gamma = 1$ , the distribution of treatment assignments is the randomization distribution, the range of the distribution of treatment assignments is the randomization distribution and the range of significance levels and Hodges–Lehmann estimates are single points. For  $\Gamma = 2$ , two subjects matched together may differ in their odds of receiving the treatment by a factor of 2, and depending on the pattern of unobserved covariates the one-sided significance level might be as large as 0.13 or as small as 0.0009, whereas the Hodges–Lehmann estimate of the decline in scores might be as small as 4.03 or as large as 11.86. The upper bound on the significance level with  $\Gamma = 2$ , namely 0.129, is obtained by using the standardized deviate

$$1.125 = (296 - 257.4)/\sqrt{1177.23}$$

derived from Table 4. The two-sided 95% confidence interval for an additive effect at  $\Gamma = 2$  is  $[-4.61, 21.36]$ . The null hypothesis of no effect becomes plausible at about  $\Gamma = 1.35$ , a fairly high degree of sensitivity to hidden bias, reflecting in part the small sample size in Table 3, the results being different in the much larger original study. Sensitivity measured by  $\Gamma$  may be compared with sensitivity in other studies; see Rosenbaum (1995), chapter 4, for a discussion of several such studies.

#### 4. Theory of separable approximation

##### 4.1. Maximizing expectations and variances yields a large random variable

Let  $Y_{Ii}$ ,  $i = 1, \dots, I$ ,  $I = 1, 2, \dots$ , be a triangular array of random variables with expectations  $E(Y_{Ii}) = \mu_{Ii}$  and finite variances  $\text{var}(Y_{Ii}) = \nu_{Ii}^2$ , where the  $I$  random variables in row  $I$  are mutually independent, such that the central limit theorem applies to the sums,  $S_I = \sum_{i=1}^I Y_{Ii}$ , so

$$\frac{1}{\sigma_I} \sum_{i=1}^I (Y_{Ii} - \mu_{Ii}) \xrightarrow{L} N(0, 1)$$

with  $\sigma_I = \sqrt{(\nu_{I1}^2 + \dots + \nu_{II}^2)}$ . Here,  $I$  is the number of matched sets and  $Y_{Ii}$  is the contribution from set  $i$  when there are  $I$  matched sets. As we move down the triangular array, we are considering larger sample sizes. A triangular array is needed because, as the sample size increases, the ranks may change for some statistics. Let  $\tilde{Y}_{Ii}$ ,  $i = 1, \dots, I$ ,  $I = 1, 2, \dots$ , be another array satisfying the same assumptions, but with  $E(\tilde{Y}_{Ii}) = \tilde{\mu}_{Ii}$ ,  $\text{var}(\tilde{Y}_{Ii}) = \tilde{\nu}_{Ii}^2$ ,  $\tilde{S}_I = \sum_{i=1}^I \tilde{Y}_{Ii}$  and

$$\frac{1}{\tilde{\sigma}_I} \sum_{i=1}^I (\tilde{Y}_{Ii} - \tilde{\mu}_{Ii}) \xrightarrow{L} N(0, 1)$$

with  $\tilde{\sigma}_I = \sqrt{(\tilde{\nu}_{I1}^2 + \dots + \tilde{\nu}_{II}^2)}$ .

Assume that  $\mu_{Ii} \geq \tilde{\mu}_{Ii}$ , and if  $\mu_{Ii} = \tilde{\mu}_{Ii}$  then  $\nu_{Ii}^2 \geq \tilde{\nu}_{Ii}^2$  for all  $i$  and  $I$ . This assumption is true by the separable algorithm in Section 3.1. Assume that there are two numbers,  $0 < \bar{\nu}^2 \leq \bar{\nu}^2 < \infty$  so that  $\bar{\nu}^2 \leq \nu_{Ii}^2 \leq \bar{\nu}^2$  and  $\bar{\nu}^2 \leq \tilde{\nu}_{Ii}^2 \leq \bar{\nu}^2$  for all  $i$  and  $I$ . Define  $\Delta = \{\mu_{Ii} - \tilde{\mu}_{Ii}; \mu_{Ii} > \tilde{\mu}_{Ii}, i = 1, \dots, I, I = 1, 2, \dots\}$ , and let  $\delta = \inf(\Delta)$ . Fix  $k > 0$  and let  $k_I = k\tilde{\sigma}_I + \sum_i \tilde{\mu}_{Ii}$  so  $\text{prob}(\tilde{S}_I \geq k_I) \rightarrow I - \Phi(k)$ .

*Proposition 1.* Under these assumptions, if  $\delta > 0$ , then for all  $\epsilon > 0$  there is an  $I^*$  such that  $\text{prob}(S_I \geq k_I) \geq \text{prob}(\tilde{S}_I \geq k_I) - \epsilon$  for  $I \geq I^*$ .

Proposition 1 says that the separable algorithm in Section 3.1 provides an approximation  $\text{prob}(S_I \geq k_I)$  to the maximum tail probability with error at most  $\epsilon$  if  $I$  is sufficiently large.

Before proving proposition 1, we make the following remarks about the assumptions. In the case of the stratified rank sum statistic, with  $n_i$  bounded, the set  $\Delta$  is a finite set, so  $\delta > 0$  is trivially true, and the variances  $\nu_{Ii}^2$  and  $\tilde{\nu}_{Ii}^2$  are uniformly bounded. As a second case, consider an aligned statistic, using the group ranks introduced by Gastwirth (1966), section 5. In a group rank statistic,  $F$  fractiles are specified,  $0 = \lambda_0 < \lambda_1 < \dots < \lambda_F < \lambda_{F+1} = 1$ , and observations with conventional ranks between  $\lfloor N\lambda_f \rfloor + 1$  and  $\lfloor N\lambda_{f+1} \rfloor$  are all assigned the same group rank  $c_f$ , where  $\lfloor x \rfloor$  is the greatest integer less or equal to  $x$ . Again  $\Delta$  is a finite set, so  $\delta > 0$  is trivially true, and the variances are uniformly bounded. Any rank statistic may be closely approximated by a group rank statistic, the approximation improving quickly as  $F$  increases.

#### 4.2. Proof of proposition 1

*Lemma 1.* Let  $a, b, x$  and  $x'$  be non-negative with  $ax + b > 0$  and  $ax' + b > 0$ . Then:

$$\sqrt{(ax' + b)} - \sqrt{(ax + b)} \leq \frac{a}{2\sqrt{(ax + b)}}(x' - x).$$

*Proof of lemma 1.* The derivative  $\partial\sqrt{(ax + b)}/\partial x = a/\{2\sqrt{(ax + b)}\}$  is decreasing in  $x$ , so  $\sqrt{(ax + b)}$  is concave. Hence the tangent at  $x$  is an overestimate, so  $\sqrt{(ax' + b)} \leq \sqrt{(ax + b)} + a(x' - x)/\{2\sqrt{(ax + b)}\}$ . □

*Proof of proposition 1.* By the central limit theorem, for all  $\epsilon > 0$  there is an  $I_1$  such that for  $I \geq I_1$  both

$$\text{prob}\{(S_I - \sum \mu_{Ii})/\sigma_I \geq k\} \geq 1 - \Phi(k) - \epsilon/2$$

and

$$\text{prob}\{(\tilde{S}_I - \sum \tilde{\mu}_{Ii})/\tilde{\sigma}_I \geq k\} = \text{prob}(\tilde{S}_I \geq k_I) \leq 1 - \Phi(k) + \epsilon/2.$$

By simple algebra,

$$\text{prob}(S_I \geq k_I) = \text{prob} \left\{ \frac{S_I - \sum_{i=1}^I \mu_{Ii}}{\sigma_I} \geq \frac{\sum_{i=1}^I (\tilde{\mu}_{Ii} - \mu_{Ii}) + k\tilde{\sigma}_I}{\sigma_I} \right\}. \tag{5}$$

If there were an  $I_2$  such that, for all  $I \geq I_2$ ,

$$k \geq \frac{\sum_{i=1}^I (\tilde{\mu}_{Ii} - \mu_{Ii}) + k\tilde{\sigma}_I}{\sigma_I},$$

then, for all  $I \geq \max(I_1, I_2)$ , probability (5) would be greater than or equal to  $1 - \Phi(k) - \epsilon/2$ , and  $\text{prob}(\tilde{S}_I \geq k_I)$  would be less than  $1 - \Phi(k) + \epsilon/2$ , so the proof would be complete. We now show that such an  $I_2$  exists. Equivalently, we show that there is an  $I_2$  such that, for all  $I \geq I_2$ ,

$$\frac{1}{I} \sum_{i=1}^I \mu_{Ii} - \tilde{\mu}_{Ii} \geq \frac{k}{\sqrt{I}} \left\{ \sqrt{\left( \frac{1}{I} \sum_{i=1}^I \tilde{\nu}_{Ii}^2 \right)} - \sqrt{\left( \frac{1}{I} \sum_{i=1}^I \nu_{Ii}^2 \right)} \right\}. \tag{6}$$

Let  $A_I = \{i: \nu_{Ii}^2 - \tilde{\nu}_{Ii}^2 < 0\}$  and let  $\pi_I = |A_I|/I$ , where the number of elements in a set  $A$  is denoted  $|A|$ . If  $i \in A_I$  then  $\mu_{Ii} - \tilde{\mu}_{Ii} \geq \delta$ , so

$$\frac{1}{I} \sum_{i=1}^I \mu_{Ii} - \tilde{\mu}_{Ii} \geq \delta\pi_I.$$

To complete the proof, it suffices to show that the right-hand side of inequality (6) is less than  $\delta\pi_I$  for all sufficiently large  $I$ . Define  $\tilde{\psi}_I$  and  $\psi_I$  to be

$$\tilde{\psi}_I = \frac{1}{I - |A_I|} \sum_{i \notin A_I} \tilde{\nu}_{Ii}^2$$

and

$$\psi_I = \frac{1}{I - |A_I|} \sum_{i \notin A_I} \nu_{Ii}^2,$$

if  $|A_I| < I$ , and to be 0 if  $|A_I| = I$ . By the definition of  $A_I$ , if  $i \notin A_I$  then  $\nu_{Ii}^2 - \tilde{\nu}_{Ii}^2 \geq 0$ , so  $0 \leq \tilde{\psi}_I \leq \psi_I$ . Also, since  $\tilde{\nu}_{Ii}^2 \leq \bar{\nu}^2$  and  $\nu_{Ii}^2 \geq \bar{\nu}^2$ ,

$$\frac{1}{I} \sum_{i \in A_I} \tilde{\nu}_{Ii}^2 \leq \pi_I \bar{\nu}^2$$

and

$$\frac{1}{I} \sum_{i \in A_I} \nu_{Ii}^2 \geq \pi_I \bar{\nu}^2.$$

It follows that

$$\frac{k}{\sqrt{I}} \left\{ \sqrt{\left( \frac{1}{I} \sum_{i=1}^I \tilde{\nu}_{Ii}^2 \right)} - \sqrt{\left( \frac{1}{I} \sum_{i=1}^I \nu_{Ii}^2 \right)} \right\} \leq \frac{k}{\sqrt{I}} [\sqrt{\{\pi_I \bar{\nu}^2 + (1 - \pi_I)\tilde{\psi}_I\}} - \sqrt{\{\pi_I \bar{\nu}^2 + (1 - \pi_I)\psi_I\}}],$$

which, by lemma 1, is less than or equal to

$$\frac{k}{\sqrt{I}} \frac{\pi_I}{2\sqrt{\{\pi_I \bar{\nu}^2 + (1 - \pi_I) \tilde{\psi}_I\}}} (\bar{\nu}^2 - \tilde{\nu}^2) \leq \frac{k}{\sqrt{I}} \frac{\pi_I}{2\tilde{\nu}} (\bar{\nu}^2 - \tilde{\nu}^2). \quad (7)$$

But the right-hand side of inequality (7) is a non-negative constant multiple of  $\pi_I/\sqrt{I}$  and so can be made less than or equal to  $\delta\pi_I$  for sufficiently large  $I$ .  $\square$

## Acknowledgements

This work was supported by grants from the National Science Foundation and the University of Pennsylvania Research Foundation.

## References

- Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996) Identification of causal effects using instrumental variables (with discussion). *J. Am. Statist. Ass.*, **91**, 444–472.
- Bross, I. D. J. (1966) Spurious effects from an extraneous variable. *J. Chron. Dis.*, **19**, 637–647.
- Copas, J. B. and Li, H. G. (1997) Inference for non-random samples (with discussion). *J. R. Statist. Soc. B*, **59**, 55–95.
- Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M. and Wynder, E. (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. *J. Natn. Cancer Inst.*, **22**, 173–203.
- Fisher, R. A. (1935) *The Design of Experiments*. Edinburgh: Oliver and Boyd.
- Gail, M., Wacholder, S. and Lubin, J. (1988) Indirect corrections for confounding under multiplicative and additive risk models. *Am. J. Indust. Med.*, **13**, 119–130.
- Gastwirth, J. L. (1966) On robust procedures. *J. Am. Statist. Ass.*, **69**, 929–948.
- (1992) Method for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables. *Jurimetrics*, **33**, 19–34.
- Gastwirth, J. L. and Nayak, T. (1997) Statistical aspects of cases concerning racial discrimination in drug sentencing: *Stephens v. State and U.S. v. Armstrong*. *J. Crim. Law Criminol.*, **87**, 583–603.
- Greenhouse, S. (1982) Jerome Cornfield's contributions to epidemiology. *Biometrics*, **28**, suppl., 33–46.
- Gu, X. S. and Rosenbaum, P. R. (1993) Comparison of multivariate matching methods: structures, distances and algorithms. *J. Comput. Graph. Statist.*, **2**, 405–420.
- Hodges, J. and Lehmann, E. (1962) Rank methods for combination of independent experiments in the analysis of variance. *Ann. Math. Statist.*, **33**, 482–497.
- (1963) Estimates of location based on rank tests. *Ann. Math. Statist.*, **34**, 598–611.
- Kempthorne, O. (1952) *The Design and Analysis of Experiments*. New York: Wiley.
- Lehmann, E. (1975) *Nonparametrics*. San Francisco: Holden-Day.
- (1986) *Testing Statistical Hypotheses*. New York: Wiley.
- Lin, D. Y., Psaty, B. M. and Kronmal, R. A. (1998) Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, **54**, 948–963.
- Rosenbaum, P. R. (1986) Dropping out of high school in the United States: an observational study. *J. Educ. Statist.*, **11**, 207–224.
- (1988) Sensitivity analysis for matching with multiple controls. *Biometrika*, **75**, 577–581.
- (1991) A characterization of optimal designs for observational studies. *J. R. Statist. Soc. B*, **53**, 597–610.
- (1993) Hodges-Lehmann point estimates of treatment effect in observational studies. *J. Am. Statist. Ass.*, **88**, 1250–1253.
- (1995) *Observational Studies*. New York: Springer.
- Rosenbaum, P. and Krieger, A. (1990) Sensitivity analysis for two-sample permutation inferences in observational studies. *J. Am. Statist. Ass.*, **85**, 493–498.
- Rosenbaum, P. and Rubin, D. (1983) Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Statist. Soc. B*, **45**, 212–218.
- Sen, P. K. (1968) On a class of aligned rank order test in two-way layouts. *Ann. Math. Statist.*, **39**, 1115–1124.
- Tardif, S. (1981) On the almost sure convergence of the permutation distribution for aligned rank test statistics in randomized block designs. *Ann. Statist.*, **9**, 190–193.